

Validation of k -Nearest Neighbor Classifiers Using Inclusion and Exclusion

Eric Bax

Yahoo Labs

BAXHOME@YAHOO.COM

Lingjie Weng

LinkedIn

LINGJIEWENG@GMAIL.COM

Xu Tian

University of California at Irvine

TIANXU03@GMAIL.COM

Editor: XXX

Abstract

This paper presents a series of PAC exponential error bounds for k -nearest neighbors classifiers, with $O(n^{-\frac{r}{2r+1}} \sqrt{k \ln n})$ error bound range for each integer $r > 0$, where n is the number of in-sample examples. This shows that k -nn classifiers, in spite of their famously fractured decision boundaries, come close to having Gaussian-style exponential error bounds with $O(n^{-\frac{1}{2}})$ bound ranges.

Keywords: Nearest neighbors, Statistical learning, Supervised learning, Generalization, Error bounds

1. Introduction

In machine learning, we begin with a set of labeled in-sample examples. We use those examples to develop a classifier, which maps from inputs to labels. The primary goal is to develop a classifier that performs well on out-of-sample data. This goal is called *generalization*. A secondary goal is to evaluate how well the classifier will perform on out-of-sample data. This is called *validation*. We do not want to sacrifice generalization for validation; we want to use all in-sample examples to develop the classifier and perform validation as well.

This paper focuses on validation of k -nearest neighbor (k -nn) classifiers. A k -nn classifier consists of the in-sample examples and a metric to determine distances between inputs. To label an input, a k -nn classifier first determines which k in-sample examples have inputs closest to the input to be classified. Then the classifier labels the input with the label shared by a majority of those k nearest neighbors. (We assume that k is odd.) We also assume binary classification, meaning that there are only two possible labels.

The error bounds used to validate classifiers in this paper are probably approximately correct (PAC) bounds. PAC bounds consist of a range and a bound on the probability that the actual out-of-sample error rate is outside the range. An effective PAC bound has a small range and a small bound failure probability. PAC error bounds include bounds based on Vapnik-Chervonenkis (VC) dimension (Vapnik and Chervonenkis, 1971), bounds for concept learning by Valiant (1984), compression-based bounds by Littlestone and Warmuth (1986), Floyd and Warmuth (1995), Blum and Langford (2003), and Bax (2008), and bounds based on worst likely assignments (Bax and Callejas, 2008).

Langford (2005) gives an overview and comparison of some types of PAC bounds for validation in machine learning.

The type of results in this paper are sometimes called conditional error bounds, because they are conditioned on a specific set of in-sample examples and hence a single classifier. There is also a history of research on the distributions of out-of-sample error rates over nearest neighbor classifiers based on random in-sample data sets, all with examples drawn i.i.d. using the same joint input-output distribution. Cover and Hart (1967) prove that the leave-one-out error rate of a classifier is an unbiased estimate of the average error rate over classifiers based on one less example than in the in-sample set. Cover (1968) shows that expected error rate converges to at most twice the optimal Bayes error rate as sample size increases. Psaltis et al. (1994) analyze how input dimension affects the rate of this convergence. For more on nearest neighbors, see the books by Devroye et al. (1996), Duda et al. (2001), and Hastie et al. (2009).

Prior research on validation of k -nn classifiers includes a method with an error bound range of $O(n^{-\frac{1}{3}})$, by Devroye and Wagner (1979), and presented in Devroye et al. (1996) (p. 415, Theorem 24.5). (We use *error bound range* to refer to twice the difference between an upper bound and the actual expected out-of-sample error rate for a classifier or to the difference between upper and lower bounds.) One way to get such a result is to use a holdout set of in-sample examples to bound the out-of-sample error rate of the classifier that is based on the remaining in-sample examples, called the holdout classifier, then bound the out-of-sample rate of disagreement between the held out classifier and the classifier based on all in-sample examples, called the full classifier. (The out-of-sample error rate of the full classifier is at most the out-of-sample error rate of the holdout classifier plus the out-of-sample rate of disagreement between the holdout and full classifiers – in the worst case, every disagreement is an error for the full classifier.) For more on this withhold-and-gap strategy, refer to Bax and Le (2015).

Let m be the number of examples withheld, and assume they are drawn uniformly at random without replacement from the n in-sample examples. The withheld examples can produce a bound on the holdout error rate with an $O(m^{-\frac{1}{2}})$ error bound range, using Hoeffding bounds (Hoeffding, 1963) or any of the other sub-Gaussian bounds on sums of independent variables (Boucheron et al., 2013). Now consider the rate of disagreement between the holdout classifier and the full classifier. The probability that at least one of the k nearest neighbors to a randomly drawn out-of-sample example is in a randomly drawn size- m holdout set is $O(\frac{m}{n})$, and this is a necessary condition for disagreement. To minimize the sum of the error bound range and the expected rate of the necessary condition for disagreement, select $m = n^{\frac{2}{3}}$. Then

$$O\left(m^{-\frac{1}{2}} + \frac{m}{n}\right) = O\left(n^{-\frac{1}{3}}\right). \quad (1)$$

A more recent method by Bax (2012) has an expected error bound range of $O(n^{-\frac{2}{5}})$. That method uses a pair of withheld data sets. The sets are used together to bound the error rate of the holdout classifier. Then each set is used to bound the difference in error rates between the holdout and full classifiers caused by withholding the examples in the other set. These bounds have ranges of $O(m^{-\frac{1}{2}})$. But we must also consider the rate of disagreement caused when both withheld data sets contain at least one of the k nearest neighbors to an out-of-sample example. This occurs with probability $O((\frac{m}{n})^2)$. Selecting $m = n^{\frac{4}{5}}$ minimizes the sum:

$$O\left(m^{-\frac{1}{2}} + \frac{m^2}{n^2}\right) = O\left(n^{-\frac{2}{5}}\right). \quad (2)$$

In this paper, we extend those error bounds by using more withheld data sets. We show that by using $r > 0$ withheld data sets, we can produce error bounds with $O(n^{-\frac{r}{2r+1}})$ expected bound range. The bounds use withheld data sets to bound rates of disagreement caused by withholding combinations of other withheld data sets, through a formula based on inclusion and exclusion. (If k is increasing with n , then use $\frac{n}{k}$ in place of n in the expected error bound sizes: $O\left(\left(\frac{n}{k}\right)^{-\frac{1}{3}}\right)$ for Devroye and Wagner (1979), $O\left(\left(\frac{n}{k}\right)^{-\frac{2}{5}}\right)$ for Bax (2012), and use $O\left(n^{-\frac{r}{2r+1}}\sqrt{k}\right)$ for this paper.)

As the number of withheld data sets grows, the number of validations for combinations of withheld data sets also grows. So using larger values of r makes sense only for larger in-sample data set sizes n . By increasing r slowly as n increases, we prove that k -nn classifiers can be validated with error bound range $O(n^{-\frac{1}{2} + \sqrt{\frac{\ln 3}{\ln n}}}\sqrt{k} \ln n)$. This result shows that k -nn classifiers come close to having Gaussian-style $O(n^{-\frac{1}{2}})$ error bounds as $n \rightarrow \infty$.

The paper is organized as follows. Section 2 outlines definitions and notation. Section 3 presents data-dependent error bounds and proofs of results about them. Section 4 extends those bounds to develop data-independent error bounds. Section 5 discusses some potential directions for future work. Appendix A shows how to modify the error bounds for use on empirical data and presents test results, and Appendix B offers an alternative approach to developing inclusion and exclusion-based error bounds.

2. Preliminaries

Let F be a set of n in-sample examples (x, y) drawn i.i.d. from a joint input-output distribution D . Inputs x are drawn from an arbitrary domain, and outputs y are drawn from $\{0, 1\}$ (binary classification). Assume there is some ordering of the examples in F , so that we may refer to examples 1 to n in F , treating F as a sequence.

Assume there is a method to measure distances between inputs. The method to compute distances need not be based on a distance metric: the method need not be symmetric and need not obey the triangle inequality. Break ties using the method from Devroye and Wagner (1979): assign each example i in F a real value Z_i drawn uniformly at random from $[0, 1]$ and do the same for each other draw x from the input space to give it a value Z . If the distance from example i in F to an x is the same as the distance from example j in F to x , then declare i to be the closer example if $|Z_i - Z| < |Z_j - Z|$ or if $|Z_i - Z| = |Z_j - Z|$ and $i < j$. Otherwise declare example j to be the closer example. This method returns the same ranking of distances to examples in F for the same input x every time the distances are measured, and it uses position within F to break a tie with probability zero.

Let g^* be the k -nn classifier based on all examples in F . To classify an input x , g^* determines the k nearest neighbors to x from F and returns the majority output over those neighbors. (Assume k is odd.) Our goal is to bound the error rate for g^* , the classifier based on a specific set of in-sample examples F . Our goal is not to bound the average error rate over random draws of in-sample data sets.

Select $r > 0$ and $m > 0$ such that $rm \leq n - k$. For each $i \in 1, \dots, r$, let validation subset V_i be the i th subset of m examples in F . For example, if $r = 2$ and $m = 1000$, then V_1 is the first thousand examples in F and V_2 is the second thousand. Let validation set $V = V_1 \cup \dots \cup V_r$. For convenience, define $R \equiv \{1, \dots, r\}$. For $S \subseteq R$, let V_S be the union of validation subsets indexed by S .

Let g_S be the k -nn classifier based on examples in

$$(F - V) \cup V_S. \quad (3)$$

In other words, g_S is based on all in-sample examples in $F - V$ and in validation subsets indexed by S . For example, $g_R \equiv g^*$.

Our PAC error bounds have probability of bound failure over draws of F . Let the subscript $F \sim D^n$ denote a probability or expectation over draws of F . We also use probabilities over out-of-sample examples (x, y) drawn i.i.d. from D and conditioned on F , which we denote by subscript $(x, y) \sim D$. For example,

$$Pr_{(x,y) \sim D} \{g^*(x) \neq y\} \equiv Pr_{(x,y) \sim D} \{g^*(x) \neq y | F\} \quad (4)$$

is the probability that g^* misclassifies an example drawn at random from D . This is also called the (conditional) out-of-sample error rate of g^* , and it is the quantity we wish to bound.

If a probability or expectation has a subscript consisting of a set, then the probability or expectation is over a uniform distribution over examples in the set indicated by the subscript, and conditioned on F . For example, using the indicator function $I(\cdot)$, defined to be one if the argument is true and zero otherwise,

$$Pr_V \{g_\emptyset(x) \neq y\} \equiv \frac{1}{|V|} \sum_{(x,y) \in V} I(g_\emptyset(x) \neq y | F) \quad (5)$$

is the average error rate of classifier g_\emptyset , which is based on the examples in $F - V$, over the validation set V . These averages are called *empirical means*.

We will use empirical means to bound out-of-sample means. If an out-of-sample mean (over $(x, y) \sim D$) is conditioned on F , but based on a set or condition that is independent of some examples in F , then the empirical mean over those examples can be used to bound the out-of-sample mean. For example, classifier g_\emptyset is based only on examples in $F - V$, so its empirical error rate over V is an unbiased estimate of its out-of-sample error rate, and the estimate is based on $|V|$ independent examples. So we can apply concentration results such as Hoeffding bounds (Hoeffding, 1963) to bound the out-of-sample error rate of g_\emptyset . This paper shows how to use inclusion and exclusion over empirical means to bridge the gap from validation of withheld classifier g_\emptyset to validation of g^* , the classifier based on all in-sample data.

Let $h(x)$ be the distance from x to its k^{th} nearest neighbor in $F - V$. Define condition $b_S(x)$ to be true if and only if $\forall i \in S$, V_i has an example closer to x than $h(x)$ and $\forall i \notin S$, V_i has no example closer to x than $h(x)$. Let condition $c_S(x)$ be true if and only if $\forall i \in S$, V_i has an example closer to x than $h(x)$. Figure 1 illustrates the definitions of $h(x)$, $b_S(x)$, and $c_S(x)$.

Let $w \leq n - rm$ and let W be the last w examples in F . Let $c'_R(x)$ be the condition that each validation subset V_i contains a nearer neighbor to x than the k^{th} nearest neighbor to x in $(F - V) - W$. Figure 2 illustrates the definition of $c'_R(x)$.

Let $B_S = \{(x, y) | b_S(x) \wedge (g_S(x) \neq y)\}$. Then B_S is the set of examples for which S indexes the validation sets that have examples closer to x than the k^{th} closest example in $F - V$ and for which the classifier based on those validation sets and on $F - V$ misclassifies x . For examples in B_S , only examples in validation sets indexed by S are close enough to x to influence the classification of the example by F , so the classifier based on those validation sets and $F - V$ must agree with the classifier based on F , meaning that $g_S(x) = g^*(x)$. Figure 3 illustrates how the error rate of the

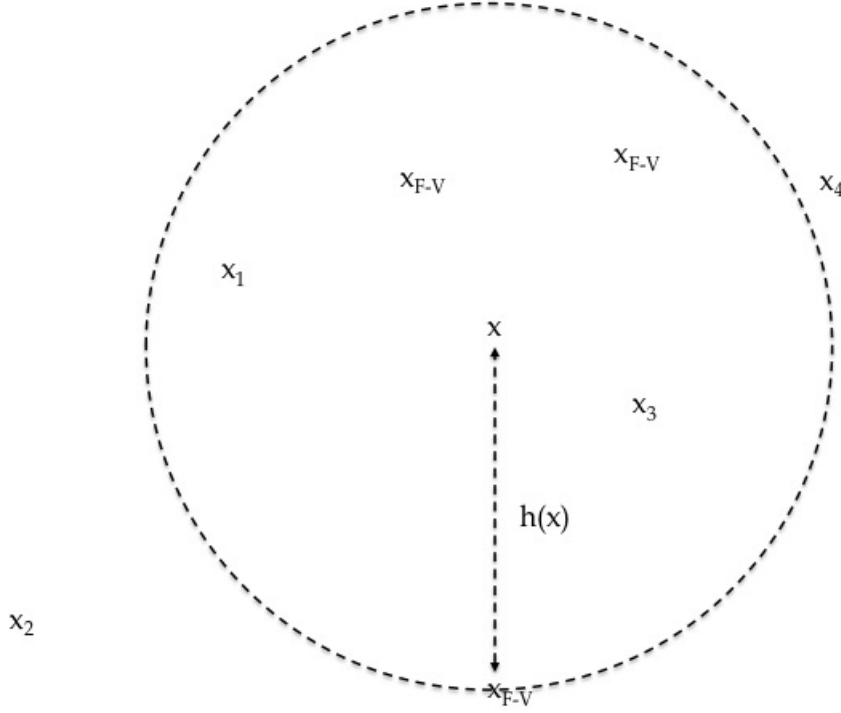


Figure 1: Suppose $k = 3$ and $r = 4$. Suppose x_1 is the closest example input to x in V_1 , x_2 is the closest to x in V_2 , x_3 is the closest to x in V_3 , and x_4 is the closest to x in V_4 . Also suppose the inputs marked x_{F-V} are the closest $k = 3$ example inputs to x in $F - V$. Then $h(x)$ is the distance from x to the third closest input marked x_{F-V} . Since x_1 and x_3 are within $h(x)$ of x , $b_{\{1,3\}}(x)$ is true, and, for all $S \neq \{1,3\}$, $b_S(x)$ is false. In contrast, $c_S(x)$ is true for all $S \subseteq \{1,3\}$, because $c_S(x)$ only requires $\forall i \in S, V_i$ has an example input closer to x than $h(x)$; for $i \notin S$, $c_S(x)$ does not depend on whether sets V_i have example inputs closer to x than $h(x)$. Since the definition of $c_S(x)$ does not depend on the examples in sets V_i for $i \notin S$, the examples in these sets (V_{R-S}) can be used to validate probabilities that are based on condition $c_S(x)$.

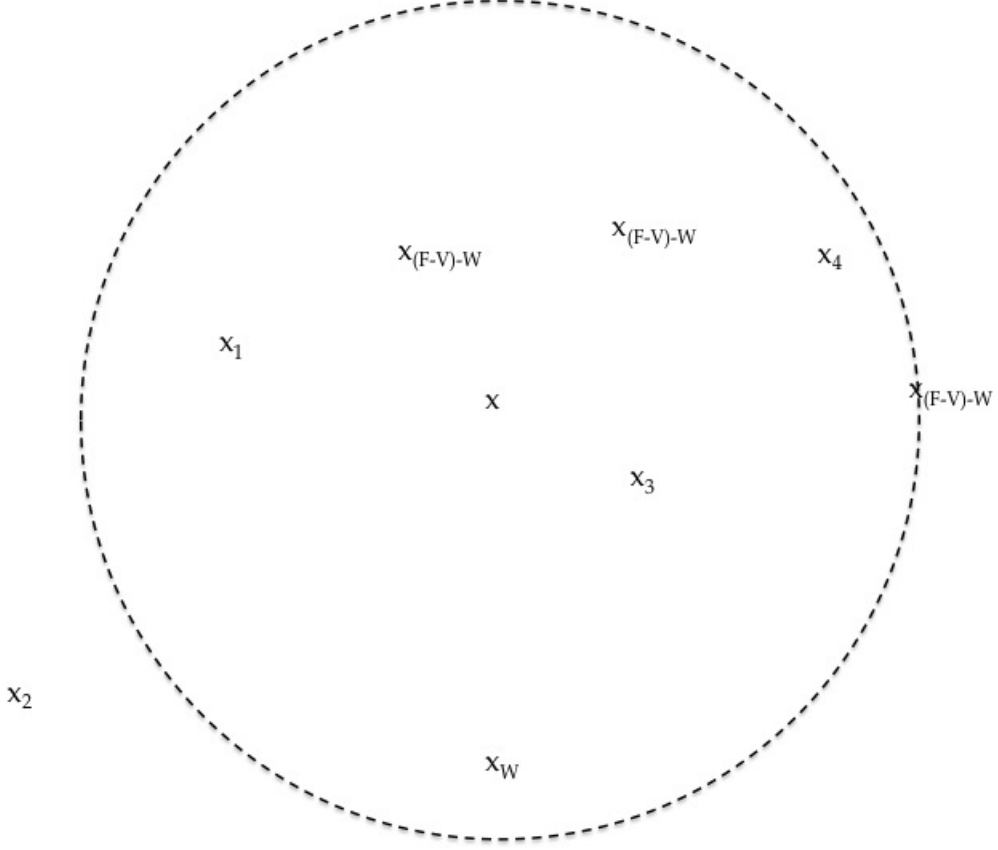


Figure 2: For condition $c_R(x)$, $S = R$, so $V_{R-S} = \emptyset$. So there are no examples in V to validate probabilities based on condition $c_R(x)$. Instead, we use a condition $c'_R(x)$ that can be validated using some examples, W , drawn from $F - V$. Condition $c'_R(x)$ requires that all validation sets V_i have an example nearer to x than the k^{th} nearest neighbor to x among the example inputs in $(F - V) - W$. Condition $c'_R(x)$ is looser than $c_R(x)$, because $c'_R(x)$ is based on the distance from x to the k^{th} closest example input in $(F - V) - W$ rather than $F - V$. As in Figure 1, suppose $k = 3$, $r = 4$, and each x_i is the nearest neighbor to x in V_i . Suppose x_W is in W , and suppose the three inputs labeled $x_{(F-V)-W}$ are the $k = 3$ nearest example inputs to x in $(F - V) - W$. Compared to Figure 1, having x_W in W increases the radius of the circle to include x_4 . However, x_2 is still outside the circle, so $c'_R(x)$ does not hold.

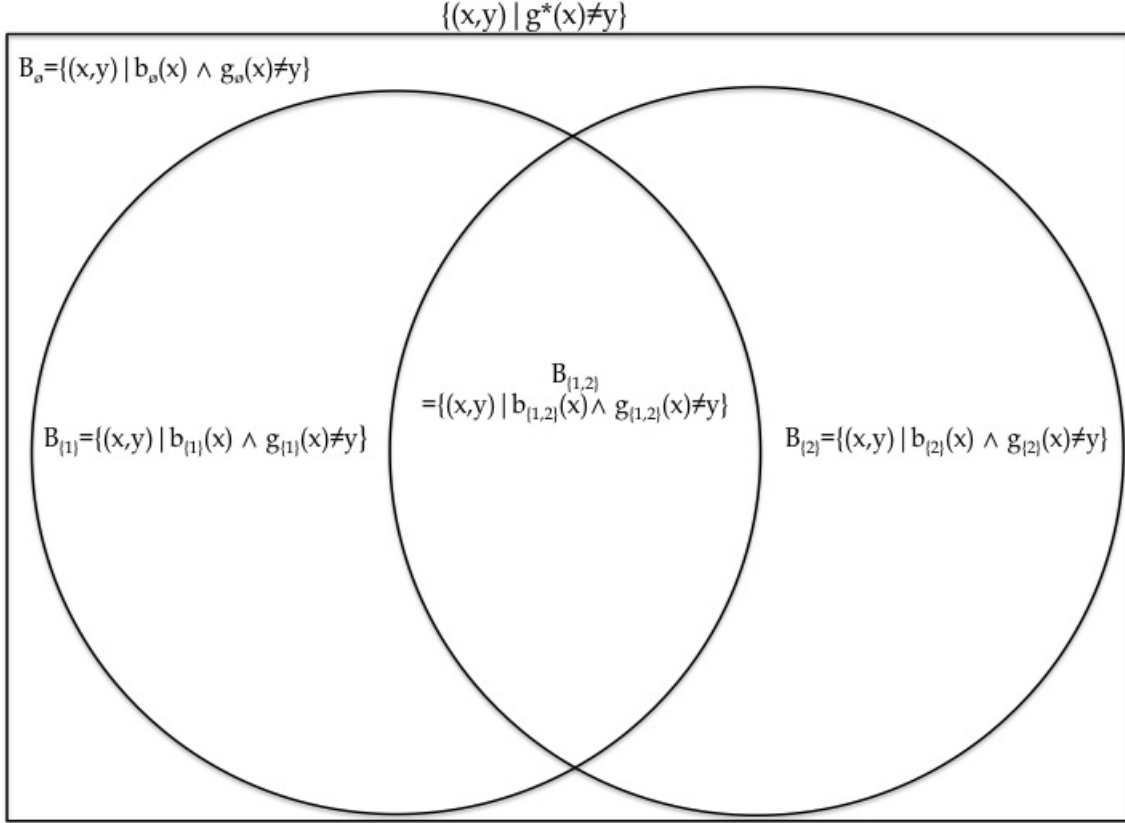


Figure 3: Set $\{(x, y) \mid g^*(x) \neq y\} = \cup_{S \subseteq R} B_S = \cup_{S \subseteq R} \{(x, y) \mid b_S(x) \wedge (g^*(x) \neq y)\}$. By definition, $B_S = \{(x, y) \mid b_S(x) \wedge (g_S(x) \neq y)\}$. But $b_S(x) \implies (g_S(x) = g^*(x))$, because $b_S(x)$ implies that validation sets not indexed by S do not contribute to the k -nn classification of x . Also, for each x , $b_S(x)$ holds for exactly one $S \subseteq R$, so sets B_S are mutually exclusive. As a result, the error rate of g^* can be decomposed: $Pr_{(x,y) \sim D} \{g^*(x) \neq y\} = \sum_{S \subseteq R} Pr_{(x,y) \sim D} \{B_S\}$. (In this figure, $r = 2$, so $R = \{1, 2\}$.)

full classifier, which is $Pr_{(x,y) \sim D} \{g^*(x) \neq y\}$, can be decomposed into a sum of $Pr_{(x,y) \sim D} \{B_S\}$ terms.

Let $C_{S,T} = \{(x, y) | c_{S \cup T}(x) \wedge (g_S(x) \neq y)\}$. Then $C_{S,T}$ is the set of examples for which validation sets indexed by $S \cup T$ all have examples closer to x than the k^{th} closest example in $F - V$ (and other validation sets may also), and the classifier based on validation sets indexed by S and on $F - V$ misclassifies x . Figure 4 illustrates how a $Pr_{(x,y) \sim D} \{B_S\}$ term can be rewritten as a signed sum of $Pr_{(x,y) \sim D} \{C_{S,T}\}$ terms, using inclusion and exclusion.

3. Data-Dependent Error Bounds

This section focuses on a method to compute data-dependent bounds on error rates of k -nn classifiers. The bounds are data-dependent in the sense that the bound range depends on the in-sample examples. The next section extends the method, to develop data-independent bounds. The data-dependent bounds require less computation, but yield results about the bound range that hold only in expectation.

The following algorithm returns a valid upper bound on the out-of-sample error rate of a k -nearest neighbor classifier with probability at least $1 - \delta - \delta_W$. The method is described in more detail in the theorems and proofs later in this section. Lines that begin with “#” are comments. The parenthetical remarks in the comments contain some terms that are defined later in this section; they can be ignored for now.

resultBound

1. inputs: data set F , $r > 0$, $|V_1|, \dots, |V_r|$, $\delta > 0$, $|W|$, $\delta_W > 0$
2. sum = 0.0.
3. # Compute an estimate for p^* and use it in a bound. (Bound t_V using s_V .)
4. Partition: $F \rightarrow (V_1, \dots, V_r, F - V)$ such that V_i is the i th $|V_i|$ examples in F , and let $V = V_1 \cup \dots \cup V_r$.
5. for $i \in \{1, \dots, r\}$:
 - (a) range = $|V_i| \sum_{S \subseteq R - \{i\}} \sum_{T \subseteq (R - \{i\}) - S} \frac{1}{|V_{R - (S \cup T)}|}$.
 - (b) values = $\left(\forall (x, y) \in V_i : |V_i| \left[\sum_{S \subseteq R - \{i\}} \sum_{T \subseteq (R - \{i\}) - S} (-1)^{|T|} \frac{1}{|V_{R - (S \cup T)}|} I((x, y) \in C_{S,T}) \right] \right)$.
 - (c) sum = sum + hoeffdingBound(values, range, $\frac{\delta}{r}$).
6. # Compute a bound on the bias of the estimate. (Bound t_W using $c'_R(x)$ values over W .)
7. Partition: $F - V \rightarrow (F - V - W, W)$ such that W is the last $|W|$ examples in F .
8. range = 1.
9. values = $(\forall (x, y) \in W : c'_R(x))$.
10. sum = sum + 2^{r-1} hoeffdingBound(values, range, δ_W).
11. return sum.

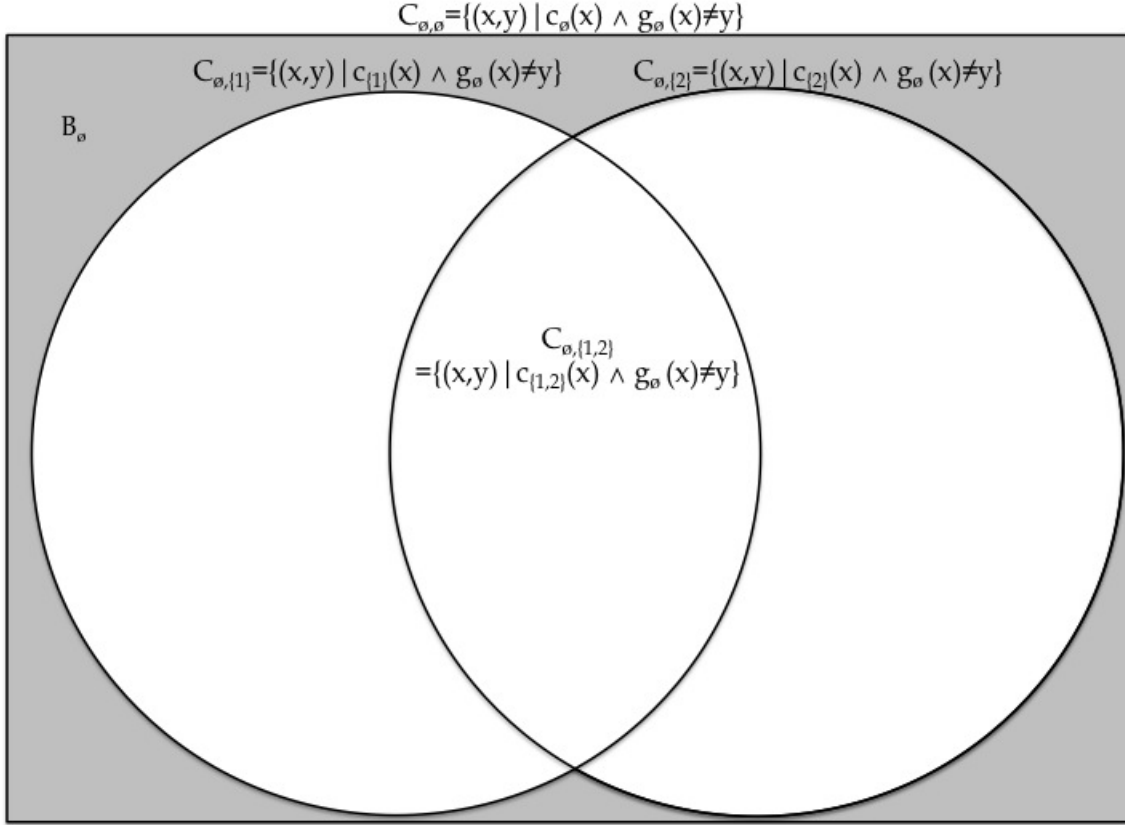


Figure 4: $Pr_{(x,y) \sim D} \{B_{\emptyset}\}$ can be rewritten as a signed sum of terms $Pr_{(x,y) \sim D} \{C_{\emptyset,T}\}$, using inclusion and exclusion. Labels directly above sets apply to those sets, except that B_{\emptyset} is the shaded area, and $C_{\emptyset,\{1,2\}}$ is the intersection. By definition, $C_{S,T} = \{(x,y) \mid c_{S \cup T}(x) \wedge (g_S(x) \neq y)\}$. Condition $c_S(x)$ requires that validation sets indexed by S have examples within $h(x)$ of x , but does not preclude other validation sets from having such examples. So $C_{\emptyset,\emptyset}$ has subsets $C_{\emptyset,\{1\}}$, $C_{\emptyset,\{2\}}$, and $C_{\emptyset,\{1,2\}}$. (Assume $r = 2$.) By adding and subtracting sets and areas, we can see that $Pr_{(x,y) \sim D} \{B_{\emptyset}\} = Pr_{(x,y) \sim D} \{C_{\emptyset,\emptyset}\} - Pr_{(x,y) \sim D} \{C_{\emptyset,\{1\}}\} - Pr_{(x,y) \sim D} \{C_{\emptyset,\{2\}}\} + Pr_{(x,y) \sim D} \{C_{\emptyset,\{1,2\}}\}$. In general, $Pr_{(x,y) \sim D} \{B_S\} = \sum_{T \subseteq R-S} (-1)^{|T|} Pr_{(x,y) \sim D} \{C_{S,T}\}$. (Note that the set system in this diagram is not the same as the set system in Figure 3.)

hoeffdingBound

1. inputs: values, range, $\delta > 0$
2. return $\text{mean}(\text{values}) + \text{range} \sqrt{\frac{\ln \frac{1}{\delta}}{2|\text{values}|}}$.

For a lower bound, change the plus signs to minus signs in line 10 of resultBound and line 2 of hoeffdingBound. We present the upper bound version of the algorithm for simplicity and because we compute upper bounds in the test section later in this paper. The results in the rest of this section are for a two-sided bound.

The fundamental result for data-dependent bounds is:

Theorem 1 For any $\delta > 0$ and $\delta_W > 0$,

$$Pr_{F \sim D^n} \{|p^* - s_V| \geq \epsilon\} \leq \delta + \delta_W, \quad (6)$$

where

$$p^* \equiv Pr_{(x,y) \sim D} \{g^*(x) \neq y\} \quad (7)$$

is the out-of-sample error rate we wish to bound,

$$s_V \equiv \sum_{S \subset R} \sum_{T \subset R-S} (-1)^{|T|} Pr_{V_{R-(S \cup T)}} \{C_{S,T}\} \quad (8)$$

is a sum of empirical means based on terms of an inclusion and exclusion formula, and

$$\epsilon \equiv r3^{r-1} \sqrt{\frac{\ln \frac{2r}{\delta}}{2m}} + 2^r \left[Pr_W \{c'_R(x)\} + \sqrt{\frac{\ln \frac{2}{\delta_W}}{2w}} \right]. \quad (9)$$

Theorem 1 gives a two-sided PAC bound on p^* :

$$p^* \in [s_V - \epsilon, s_V + \epsilon] \text{ with probability at least } 1 - (\delta + \delta_W). \quad (10)$$

Refer to ϵ from Theorem 1 as the error bound range. Refer to $E_{F \sim D^n} \{\epsilon\}$ as the expected error bound range. If we hold r constant, then increasing m decreases the first term of ϵ from the RHS of Equation 9, which tightens the error bound. But it also tends to increase $Pr_W \{c'_R(x)\}$, because larger validation sets make it more likely that validation sets will have some of the closest neighbors to examples in W . This loosens the bound. Selecting m to mediate this tradeoff yields:

Theorem 2 For constant r and appropriate choices of m and w (which are detailed in the proof),

$$E_{F \sim D^n} \{\epsilon\} = O(n^{-\frac{r}{2r+1}} \sqrt{k}). \quad (11)$$

Now suppose we allow r to grow with n . Then we can show:

Corollary 3 For a method to increase r as n increases,

$$E_{F \sim D^n} \{\epsilon\} = O\left(n^{-\frac{1}{2} + \sqrt{\frac{\ln 3}{\ln n}}} \sqrt{k \ln n} \sqrt{\ln \sqrt{\ln n}}\right). \quad (12)$$

Now we prove these results.

Proof [of Theorem 1] Recall that $h(x)$ is the distance from x to its k^{th} nearest neighbor in $F - V$. Also, recall that condition $b_S(x)$ is true if and only if $\forall i \in S$, V_i has an example closer to x than $h(x)$ and $\forall i \notin S$, V_i has no example closer to x than $h(x)$. As a result, for each x , $b_S(x)$ holds for exactly one S , and

$$b_S(x) \implies (g_S(x) = g^*(x)), \quad (13)$$

because the k nearest neighbors to x are in $F - V$ or validation subsets indexed by S . So

$$Pr_{(x,y) \sim D} \{g^*(x) \neq y\} = \sum_{S \subseteq R} Pr_{(x,y) \sim D} \{b_S(x) \wedge (g_S(x) \neq y)\} = \sum_{S \subseteq R} Pr_{(x,y) \sim D} \{B_S\}. \quad (14)$$

Recall that condition $c_S(x)$ holds if and only if $\forall i \in S$, V_i has an example closer to x than $h(x)$. Note that c_S is a looser condition than b_S , because c_S does not require that $\forall i \notin S$, V_i has no example closer to x than $h(x)$. So, for each x , $c_S(x)$ may hold for multiple S .

The following lemma states the out-of-sample error rate in terms of probabilities based on conditions c_S . We want terms based on conditions c_S rather than b_S because we can validate terms based on c_S , as we show later. The lemma is derived by applying inclusion and exclusion to each $Pr_{(x,y) \sim D} \{B_S\}$ term in the RHS of Equation 14, to rewrite the term as a sum of signed $Pr_{(x,y) \sim D} \{C_{S,T}\}$ terms.

Lemma 4

$$Pr_{(x,y) \sim D} \{g^*(x) \neq y\} = \sum_{S \subseteq R} \sum_{T \subseteq R-S} (-1)^{|T|} Pr_{(x,y) \sim D} \{C_{S,T}\}. \quad (15)$$

Proof [of Lemma 4] To prove this lemma, we will show:

$$\forall S \subseteq R : Pr_{(x,y) \sim D} \{B_S\} = \sum_{T \subseteq R-S} (-1)^{|T|} Pr_{(x,y) \sim D} \{C_{S,T}\}. \quad (16)$$

Note that

$$b_S(x) = c_S(x) \wedge \neg \left[\bigvee_{i \in R-S} c_{S \cup \{i\}}(x) \right], \quad (17)$$

because b_S requires that $\forall i \notin R - S$, V_i has no examples closer to x than $h(x)$. Similarly,

$$(b_S(x) \wedge (g_S(x) \neq y)) \quad (18)$$

$$= (c_S(x) \wedge (g_S(x) \neq y)) \quad (19)$$

$$\wedge \neg \left[\bigvee_{i \in R-S} (c_{S \cup \{i\}}(x) \wedge (g_S(x) \neq y)) \right]. \quad (20)$$

In terms of $B_S = \{(x, y) | b_S(x) \wedge (g_S(x) \neq y)\}$ and $C_{S,T} = \{(x, y) | c_{S \cup T}(x) \wedge (g_S(x) \neq y)\}$, this is:

$$B_S = C_{S, \emptyset} - \bigcup_{i \in \{R-S\}} C_{S, \{i\}}. \quad (21)$$

So

$$Pr_{(x,y) \sim D} \{B_S\} = Pr_{(x,y) \sim D} \{C_{S, \emptyset}\} - Pr_{(x,y) \sim D} \left\{ \bigcup_{i \in \{R-S\}} C_{S, \{i\}} \right\} \quad (22)$$

$$+ Pr_{(x,y) \sim D} \left\{ \bigcup_{i \in \{R-S\}} C_{S, \{i\}} - C_{S, \emptyset} \right\}. \quad (23)$$

Note that $\forall i : C_{S,\{i\}} \subseteq C_{S,\emptyset}$. So the term $Pr_{(x,y) \sim D} \{ \cup_{i \in \{R-S\}} C_{S,\{i\}} - C_{S,\emptyset} \}$ is zero:

$$Pr_{(x,y) \sim D} \{B_S\} = Pr_{(x,y) \sim D} \{C_{S,\emptyset}\} - Pr_{(x,y) \sim D} \{ \cup_{i \in \{R-S\}} C_{S,\{i\}} \}. \quad (24)$$

Apply inclusion and exclusion to the probability of a union:

$$Pr_{(x,y) \sim D} \{B_S\} \quad (25)$$

$$= Pr_{(x,y) \sim D} \{C_{S,\emptyset}\} \quad (26)$$

$$- \sum_{i_1 \in R-S} Pr_{(x,y) \sim D} \{C_{S,\{i_1\}}\} \quad (27)$$

$$+ \sum_{\{i_1, i_2\} \subseteq R-S} Pr_{(x,y) \sim D} \{C_{S,\{i_1\}} \cap C_{S,\{i_2\}}\} \quad (28)$$

$$- \sum_{\{i_1, i_2, i_3\} \subseteq R-S} Pr_{(x,y) \sim D} \{C_{S,\{i_1\}} \cap C_{S,\{i_2\}} \cap C_{S,\{i_3\}}\} \quad (29)$$

$$\pm \dots \quad (30)$$

By the definition of $C_{S,T}$,

$$C_{S,\{i_1\}} \cap \dots \cap C_{S,\{i_j\}} = C_{S,\{i_1, \dots, i_j\}}. \quad (31)$$

So

$$Pr_{(x,y) \sim D} \{B_S\} \quad (32)$$

$$= Pr_{(x,y) \sim D} \{C_{S,\emptyset}\} \quad (33)$$

$$- \sum_{i_1 \in R-S} Pr_{(x,y) \sim D} \{C_{S,\{i_1\}}\} \quad (34)$$

$$+ \sum_{\{i_1, i_2\} \subseteq R-S} Pr_{(x,y) \sim D} \{C_{S,\{i_1, i_2\}}\} \quad (35)$$

$$- \sum_{\{i_1, i_2, i_3\} \subseteq R-S} Pr_{(x,y) \sim D} \{C_{S,\{i_1, i_2, i_3\}}\} \quad (36)$$

$$\pm \dots \quad (37)$$

Equivalently,

$$Pr_{(x,y) \sim D} \{B_S\} = \sum_{T \subseteq R-S} (-1)^{|T|} Pr_{(x,y) \sim D} \{C_{S,T}\}. \quad (38)$$

Substitute this expression for each term in the RHS of Equation 14 to prove the lemma. ■

Separate the RHS of Equation 15 into terms for which $S \cup T \subset R$:

$$t_V = \sum_{S \subset R} \sum_{T \subset R-S} (-1)^{|T|} Pr_{(x,y) \sim D} \{C_{S,T}\} \quad (39)$$

and terms for which $S \cup T = R$:

$$t_W = \sum_{S \subseteq R} (-1)^{|R-S|} Pr_{(x,y) \sim D} \{C_{S,R-S}\}. \quad (40)$$

We will use empirical means over validation sets V_i to bound t_V . Then we will bound t_W using an empirical mean over W .

Rewrite t_V by gathering terms for each $i \in R$ that have $i \notin S \cup T$ and multiplying each term by $\frac{|V_i|}{|V_{R-(S \cup T)}|}$, so that the sum of these coefficients for each term is one.

$$t_V = \sum_{i=1}^r \left[\sum_{S \subseteq R-\{i\}} \sum_{T \subseteq (R-\{i\})-S} \frac{|V_i|}{|V_{R-(S \cup T)}|} (-1)^{|T|} Pr_{(x,y) \sim D} \{C_{S,T}\} \right]. \quad (41)$$

Convert the probability to the expectation of an indicator function, and use the linearity of expectation:

$$t_V = \sum_{i=1}^r E_{(x,y) \sim D} \left\{ \sum_{S \subseteq R-\{i\}} \sum_{T \subseteq (R-\{i\})-S} \frac{|V_i|}{|V_{R-(S \cup T)}|} (-1)^{|T|} I((x,y) \in C_{S,T}) \right\}. \quad (42)$$

Define

$$f_i(x,y) \equiv \sum_{S \subseteq R-\{i\}} \sum_{T \subseteq (R-\{i\})-S} \frac{|V_i|}{|V_{R-(S \cup T)}|} (-1)^{|T|} I((x,y) \in C_{S,T}). \quad (43)$$

Then

$$t_V = \sum_{i=1}^r E_{(x,y) \sim D} \{f_i(x,y)\}. \quad (44)$$

Note that each $f_i(x,y)$ is independent of the examples in V_i , since the sums in Equation 43 are over sets S and T that lack i , and $i \notin S \cup T$ implies

$$\forall (x,y) : I((x,y) \in C_{S,T})|F = I((x,y) \in C_{S,T})|(F - V_i). \quad (45)$$

So for each $i \in R$, we can use the empirical mean $E_{V_i} \{f_i(x,y)\}$ to bound $E_{(x,y) \sim D} \{f_i(x,y)\}$.

To apply the Hoeffding Inequality (Hoeffding, 1963), we need to know the length of the range of $f_i(x,y)$. Each term in $f_i(x,y)$ has a range of length at most one, since $|V_i| \leq |V_{R-(S \cup T)}|$. There are as many terms as there ways to partition $R - \{i\}$ into three subsets: S , T , and $R - (S \cup T)$. So there are 3^{r-1} terms. So $f_i(x,y)$ has range length at most 3^{r-1} .

Recall that $|V_i| = m$. Apply the Hoeffding Inequality with $\frac{\delta}{r}$ in place of δ . Then

$$\forall i \in R : Pr_{F \sim D^n} \left\{ |E_{V_i} \{f_i(x,y)\} - E_{(x,y) \sim D} \{f_i(x,y)\}| \geq 3^{r-1} \sqrt{\frac{\ln \frac{2r}{\delta}}{2m}} \right\} \leq \frac{\delta}{r}. \quad (46)$$

Using the sum bound on the union of these probabilities:

$$Pr_{F \sim D^n} \left\{ \left| \sum_{i=1}^r E_{V_i} \{f_i(x,y)\} - \sum_{i=1}^r E_{(x,y) \sim D} \{f_i(x,y)\} \right| \geq r 3^{r-1} \sqrt{\frac{\ln \frac{2r}{\delta}}{2m}} \right\} \leq \delta. \quad (47)$$

Note that

$$\sum_{i=1}^r E_{V_i} \{f_i(x, y)\} \quad (48)$$

$$= \sum_{i=1}^r \sum_{(x,y) \in V_i} \frac{1}{|V_i|} \left[\sum_{S \subseteq R - \{i\}} \sum_{T \subseteq (R - \{i\}) - S} \frac{|V_i|}{|V_{R-(S \cup T)}|} (-1)^{|T|} I((x, y) \in C_{S,T}) \right] \quad (49)$$

$$= \sum_{i=1}^r \sum_{(x,y) \in V_i} \left[\sum_{S \subseteq R - \{i\}} \sum_{T \subseteq (R - \{i\}) - S} \frac{1}{|V_{R-(S \cup T)}|} (-1)^{|T|} I((x, y) \in C_{S,T}) \right] \quad (50)$$

$$= \sum_{S \subseteq R} \sum_{T \subseteq R - S} (-1)^{|T|} Pr_{V_{R-(S \cup T)}} \{C_{S,T}\} \quad (51)$$

$$= s_V, \quad (52)$$

as defined in the statement of Theorem 1. Substitute s_V and Equation 44 into Inequality 47:

$$Pr_{F \sim D^n} \left\{ |s_V - t_V| \geq r 3^{r-1} \sqrt{\frac{\ln \frac{2r}{\delta}}{2m}} \right\} \leq \delta. \quad (53)$$

(We can get a similar result by applying the Hoeffding Inequality to each sum of terms that have the same set of validation data, $V_{R-(S \cup T)}$, then applying a union bound. See the appendix for details.)

Now consider t_W :

$$t_W = \sum_{S \subseteq R} (-1)^{|R-S|} Pr_{(x,y) \sim D} \{C_{S,R-S}\} \quad (54)$$

$$= \sum_{S \subseteq R} (-1)^{|R-S|} Pr_{(x,y) \sim D} \{c_R(x) \wedge (g_S(x) \neq y)\}. \quad (55)$$

Note that

$$t_W \in [-2^{r-1} Pr_{(x,y) \sim D} \{c_R(x)\}, 2^{r-1} Pr_{(x,y) \sim D} \{c_R(x)\}]. \quad (56)$$

To estimate $Pr_{(x,y) \sim D} \{c_R(x)\}$, select a sample size $w \leq n - rm$ and let sample W be the last w examples in $F - V$. Let $c'_R(x)$ be the condition that each validation subset V_i contains a nearer neighbor to x than the k^{th} nearest neighbor to x in $(F - V) - W$. Since $(F - V) - W \subset F - V$, $c_R(x)$ implies $c'_R(x)$. So

$$Pr_{(x,y) \sim D} \{c'_R(x)\} \geq Pr_{(x,y) \sim D} \{c_R(x)\}. \quad (57)$$

Hence

$$t_W \in [-2^{r-1} Pr_{(x,y) \sim D} \{c'_R(x)\}, 2^{r-1} Pr_{(x,y) \sim D} \{c'_R(x)\}]. \quad (58)$$

Use empirical mean $Pr_W \{c'_R(x)\}$ to estimate $Pr_{(x,y) \sim D} \{c'_R(x)\}$. Let

$$\epsilon_W = Pr_{(x,y) \sim D} \{c'_R(x)\} - Pr_W \{c'_R(x)\}. \quad (59)$$

Again using the Hoeffding Inequality, for any $\delta_W > 0$,

$$Pr_{F \sim D^n} \left\{ |\epsilon_W| \geq \sqrt{\frac{\ln \frac{2}{\delta_W}}{2w}} \right\} \leq \delta_W. \quad (60)$$

So

$$Pr_{F \sim D^n} \left\{ |t_W| \geq 2^r \left[Pr_W \{c'_R(x)\} + \sqrt{\frac{\ln \frac{2}{\delta_W}}{2w}} \right] \right\} \leq \delta_W. \quad (61)$$

Combining this with the bound for t_V from Inequality 53, the probability (over random draws of F) that the absolute value of the difference between the out-of-sample error rate of g^* and the estimate of t_V using empirical means exceeds

$$\epsilon \equiv r3^{r-1} \sqrt{\frac{\ln \frac{2r}{\delta}}{2m}} + 2^r \left[Pr_W \{c'_R(x)\} + \sqrt{\frac{\ln \frac{2}{\delta_W}}{2w}} \right] \quad (62)$$

is at most $\delta + \delta_W$, which completes the proof of Theorem 1. \blacksquare

Proof [of Theorem 2] Our goal is to bound the expected value of ϵ , the range for the two-sided bound, from Equation 62. To do that, we prove the following lemma about the expected value of $Pr_W \{c'_R(x)\}$.

Lemma 5

$$E_{F \sim D^n} \{Pr_W \{c'_R(x)\}\} \leq \left(\frac{(k+r-1)m}{n-w} \right)^r e^r. \quad (63)$$

Proof [of Lemma 5] Define $c''_R(x)$ to be the condition that the $k+r-1$ nearest neighbors to x in $F-W$ include at least r examples from V . Condition $c''_R(x)$ is a necessary condition for $c'_R(x)$, because otherwise the k^{th} nearest neighbor to x in $F-W$ is closer to x than the nearest neighbor from at least one of V_1, \dots, V_r . So

$$E_{F \sim D^n} \{Pr_W \{c''_R(x)\}\} \geq E_{F \sim D^n} \{Pr_W \{c'_R(x)\}\}. \quad (64)$$

For each x , for each draw of F , every permutation of $1, \dots, n-w$ is equally likely to be the ranking by distance to x for the indices of examples in $F-W$. So

$$\forall x : Pr_{F \sim D^n} \{c''_R(x)\} \quad (65)$$

is the same as the probability that the first $k+r-1$ samples drawn uniformly without replacement from $\{1, \dots, n-w\}$ contain at least r elements from $\{1, \dots, |V|\}$. So it is the tail of a hypergeometric distribution:

$$\forall x : Pr_{F \sim D^n} \{c''_R(x)\} = \sum_{i=r}^{k+r-1} \frac{\binom{k+r-1}{i} \binom{(n-w)-(k+r-1)}{rm-i}}{\binom{n-w}{rm}}. \quad (66)$$

Using a hypergeometric tail bound from Chvátal (1979), this is

$$\leq \left(\frac{(k+r-1)m}{n-w} \right)^r \left[\left(1 + \frac{1}{m-1} \right) \left(1 - \frac{k+r-1}{n-w} \right) \right]^{(m-1)r} \quad (67)$$

$$\leq \left(\frac{(k+r-1)m}{n-w} \right)^r \left[\left(1 + \frac{1}{m-1} \right)^{m-1} \right]^r \quad (68)$$

$$\leq \left(\frac{(k+r-1)m}{n-w} \right)^r e^r. \quad (69)$$

Since the examples in W are independent of those in $F - W$,

$$E_{F \sim D^n} \{Pr_W \{c_R''(x)\}\} = Pr_{F \sim D^n} \{c_R''(x)\}. \quad (70)$$

■

Lemma 5 implies

$$E_{F \sim D^n} \{\epsilon\} \leq r3^{r-1} \sqrt{\frac{\ln \frac{2r}{\delta}}{2m}} + 2^r \left[\left(\frac{(k+r-1)m}{n-w} \right)^r e^r + \sqrt{\frac{\ln \frac{2}{\delta_W}}{2w}} \right]. \quad (71)$$

Let $w = m$. Let

$$m = \frac{(n-m)^{\frac{r}{r+1}}}{(k+r-1)e}. \quad (72)$$

(In practice, use the nearest integer to the solution for m .) Then

$$E_{F \sim D^n} \{\epsilon\} \leq (n-m)^{-\frac{r}{2r+1}} \left[r3^{r-1} \sqrt{\frac{1}{2}(k+r-1)e \ln \frac{2r}{\delta}} + 2^r \left(1 + \sqrt{\frac{1}{2}(k+r-1)e \ln \frac{2}{\delta_W}} \right) \right]. \quad (73)$$

Treating r as a constant,

$$E_{F \sim D^n} \{\epsilon\} = O(n^{-\frac{r}{2r+1}} \sqrt{k}), \quad (74)$$

which completes the proof of Theorem 2. ■

Proof [of Corollary 3] Suppose we do not hold r constant, and instead increase r slowly as n increases. Then, based on Inequality 73,

$$E_{F \sim D^n} \{\epsilon\} = O\left(n^{-\frac{r}{2r+1}} 3^{r-1} r \sqrt{k} \sqrt{\ln r}\right). \quad (75)$$

Let $r = \lceil C\sqrt{\ln n} \rceil$. Then

$$E_{F \sim D^n} \{\epsilon\} = O\left(n^{-\frac{\lceil C\sqrt{\ln n} \rceil}{2\lceil C\sqrt{\ln n} \rceil+1}} 3^{C\sqrt{\ln n}} \sqrt{k \ln n} \sqrt{\ln \sqrt{\ln n}}\right). \quad (76)$$

Expand the fraction in the exponent on n and convert the exponent on 3 to an exponent on n :

$$E_{F \sim D^n} \{\epsilon\} = O\left(n^{-\frac{1}{2} + \frac{1}{4\lceil C\sqrt{\ln n} \rceil+2} + \frac{C\ln 3}{\sqrt{\ln n}}} \sqrt{k \ln n} \sqrt{\ln \sqrt{\ln n}}\right). \quad (77)$$

Let $C = \frac{1}{2\sqrt{\ln 3}}$ and remove the 2 from the denominator of the exponent to simplify. Then

$$E_{F \sim D^n} \{\epsilon\} = O\left(n^{-\frac{1}{2} + \sqrt{\frac{\ln 3}{\ln n}}} \sqrt{k \ln n} \sqrt{\ln \sqrt{\ln n}}\right). \quad (78)$$

■

4. Data-Independent Error Bounds

This section extends the data-dependent error bounds from the last section to develop data-independent error bounds. These bounds require more computation and are slightly weaker on average, but they enable us to prove that there are exponential PAC error bounds for k -nn classifiers with error bound range, rather than just expected error bound range, of $O\left(n^{-\frac{1}{2} + \sqrt{\frac{2}{\ln n}}}\right)$.

Let P be the set of all permutations of $1, \dots, n$. Let Q be a size- q random sample of permutations drawn uniformly with replacement from P . For permutation σ , let σF be the in-sample examples F permuted according to σ , so that the i th example in σF is the example in the position of F indexed by the i th element of σ . Also, for $i \in R$, let $V_i(\sigma)$ be the i th subset of m examples in σF , and let $V(\sigma) = V_1(\sigma) \cup \dots \cup V_r(\sigma)$.

Let $t_V(\sigma)$ be t_V , but with σF in place of F and hence $V_i(\sigma)$ in place of V_i for all $i \in R$, $V(\sigma)$ in place of V , and $\sigma F - V(\sigma)$ in place of $F - V$. Similarly, let $s_V(\sigma)$ be s_V , but with σF in place of F .

We will show:

Theorem 6 For $\delta > 0$ and $\delta_q > 0$:

$$Pr_{F \sim D^n} \{|p^* - E_Q \{s_V(\sigma)\}| \geq \epsilon_q\} \leq \delta + \delta_q, \quad (79)$$

where

$$\epsilon_q = r3^{r-1} \sqrt{\frac{\ln \frac{2rq}{\delta}}{2m}} + 2^r \left[\sqrt{\frac{\ln \frac{2}{\delta_q}}{2q}} + \left(\frac{(k+r-1)m}{n} \right)^r e^r \right]. \quad (80)$$

Also, for choices of m and q detailed in the proof:

Corollary 7 Holding r constant, ϵ_q is $O\left(n^{-\frac{r}{2r+1}} \sqrt{k \ln n}\right)$.

If r is allowed to grow with n , then we can choose r such that:

Corollary 8

$$\epsilon_q = O\left(n^{-\frac{1}{2} + \sqrt{\frac{\ln 3}{\ln n}}} \sqrt{k \ln n}\right). \quad (81)$$

Proof [of Theorem 6] Similar to $t_V(\sigma)$, let $t_W(\sigma)$ be t_W , but with σF in place of F . Also, let $p^*(\sigma)$ be the error rate of the k -nn classifier based on σF . Recall that

$$p^* = t_V + t_W. \quad (82)$$

Substitute σF for F :

$$\forall \sigma \in P : p^*(\sigma) = t_V(\sigma) + t_W(\sigma). \quad (83)$$

For all $\sigma \in P$, $p^*(\sigma) = p^*$, because the k -nn classifier based on σF is based on the same examples as the k -nn classifiers based on F , and the probability of example ordering in F or σF being used for distance tie-breaking is zero. So

$$\forall \sigma \in P : p^* = t_V(\sigma) + t_W(\sigma). \quad (84)$$

Average over the sample permutations in Q :

$$E_Q \{p^*\} = E_Q \{t_V(\sigma)\} + E_Q \{t_W(\sigma)\}. \quad (85)$$

Since p^* does not depend on $\sigma \in Q$:

$$p^* = E_Q \{t_V(\sigma)\} + E_Q \{t_W(\sigma)\}. \quad (86)$$

We will use an estimate to bound $E_Q \{t_V(\sigma)\}$, then we will bound $E_Q \{t_W(\sigma)\}$. For $E_Q \{t_V(\sigma)\}$, let $s_V(\sigma)$ be s_V , but with σF in place of F . Use Equation 53, and substitute $\frac{\delta}{q}$ for δ .

$$\forall \sigma \in Q : Pr_{F \sim D^n} \left\{ |s_V(\sigma) - t_V(\sigma)| \geq r3^{r-1} \sqrt{\frac{\ln \frac{2rq}{\delta}}{2m}} \right\} \leq \frac{\delta}{q}. \quad (87)$$

Use the sum bound for the probability of a union:

$$Pr_{F \sim D^n} \left\{ \exists \sigma \in Q : |s_V(\sigma) - t_V(\sigma)| \geq r3^{r-1} \sqrt{\frac{\ln \frac{2rq}{\delta}}{2m}} \right\} \leq \delta. \quad (88)$$

Multiply both sides of the inner inequality by $\frac{1}{q}$:

$$Pr_{F \sim D^n} \left\{ \exists \sigma \in Q : \frac{1}{q} |s_V(\sigma) - t_V(\sigma)| \geq \frac{1}{q} r3^{r-1} \sqrt{\frac{\ln \frac{2rq}{\delta}}{2m}} \right\} \leq \delta. \quad (89)$$

Sum the inner inequalities over $\sigma \in Q$, and move $\frac{1}{q}$ and the summation over $\sigma \in Q$ inside the absolute values:

$$Pr_{F \sim D^n} \left\{ \left| \sum_{\sigma \in Q} \frac{1}{q} s_V(\sigma) - \sum_{\sigma \in Q} \frac{1}{q} t_V(\sigma) \right| \geq r3^{r-1} \sqrt{\frac{\ln \frac{2rq}{\delta}}{2m}} \right\} \leq \delta. \quad (90)$$

So we have:

$$Pr_{F \sim D^n} \left\{ |E_Q \{s_V(\sigma)\} - E_Q \{t_V(\sigma)\}| \geq r3^{r-1} \sqrt{\frac{\ln \frac{2rq}{\delta}}{2m}} \right\} \leq \delta. \quad (91)$$

This is a bound on $E_Q \{t_V(\sigma)\}$, the first term on the RHS of Equation 86.

Now consider how to bound $E_Q \{t_W(\sigma)\}$, the second term on the RHS of Equation 86. Let

$$E_{\sigma \sim P} \{t_W(\sigma)\} \equiv \sum_{\sigma \in P} \frac{1}{|P|} t_W(\sigma) \quad (92)$$

be the mean of $t_W(\sigma)$ over permutations drawn uniformly at random from P . Note that $E_Q \{t_W(\sigma)\}$ is a sample mean for the distribution mean $E_{\sigma \sim P} \{t_W(\sigma)\}$. Think of $E_Q \{t_W(\sigma)\}$ as an empirical mean that might be used to bound $E_{\sigma \sim P} \{t_W(\sigma)\}$ through a concentration inequality. However,

we will use the concentration inequality “backward”, using a bound on $E_{\sigma \sim P} \{t_W(\sigma)\}$ to bound $E_Q \{t_W(\sigma)\}$.

By Equation 56,

$$\forall \sigma \in Q : t_W(\sigma) \in [-2^{r-1}, 2^{r-1}]. \quad (93)$$

So, by applying a Hoeffding bound (Hoeffding, 1963), for $\delta_q > 0$:

$$Pr_{F \sim D^n} \left\{ |E_{\sigma \sim P} \{t_W(\sigma)\} - E_Q \{t_W(\sigma)\}| \geq 2^r \sqrt{\frac{\ln \frac{2}{\delta_q}}{2q}} \right\} \leq \delta_q. \quad (94)$$

So

$$Pr_{F \sim D^n} \left\{ |E_Q \{t_W(\sigma)\}| \geq 2^r \sqrt{\frac{\ln \frac{2}{\delta_q}}{2q}} + |E_{\sigma \sim P} \{t_W(\sigma)\}| \right\} \leq \delta_q. \quad (95)$$

Let $c_R(x, \sigma)$ be $c_R(x)$, but with σF in place of F . By Equation 56,

$$\forall \sigma \in P : |t_W(\sigma)| \leq 2^{r-1} Pr_{(x,y) \sim D} \{c_R(x, \sigma)\}. \quad (96)$$

So

$$|E_{\sigma \sim P} \{t_W(\sigma)\}| \leq E_{\sigma \sim P} \{|t_W(\sigma)|\} \leq 2^{r-1} E_{\sigma \sim P} \{Pr_{(x,y) \sim D} \{c_R(x, \sigma)\}\}. \quad (97)$$

Condition $c_R(x, \sigma)$ is that the first through r th subsets of m examples from σF each contribute a nearer neighbor to x than the k th nearest neighbor from the remaining examples in σF . Define $c_R'''(x)$ to be the condition that the nearest $k+r-1$ neighbors to x in σF include at least r examples from the first rm examples in σF . Note that $c_R'''(x, \sigma) \implies c_R(x, \sigma)$. So

$$\forall \sigma \in P, x : Pr_{(x,y) \sim D} \{c_R'''(x, \sigma)\} \geq Pr_{(x,y) \sim D} \{c_R(x, \sigma)\}. \quad (98)$$

Substitute into the RHS of Inequality 97:

$$|E_{\sigma \sim P} \{t_W(\sigma)\}| \leq 2^{r-1} E_{\sigma \sim P} \{Pr_{(x,y) \sim D} \{c_R'''(x, \sigma)\}\}. \quad (99)$$

By the linearity of expectation:

$$E_{\sigma \sim P} \{Pr_{(x,y) \sim D} \{c_R'''(x, \sigma)\}\} = E_{(x,y) \sim D} \{Pr_{\sigma \sim P} \{c_R'''(x, \sigma)\}\}. \quad (100)$$

For each x , for each draw of σ , every permutation of $1, \dots, n$ is equally likely to be the ranking by distance to x of the indices of examples in σF . So

$$\forall x : Pr_{\sigma \sim P} \{c_R'''(x)\} \quad (101)$$

is the same as the probability that the first $k+r-1$ samples drawn uniformly without replacement from $\{1, \dots, n\}$ contain at least r elements from $\{1, \dots, rm\}$. It is the tail of a hypergeometric distribution. So, using Expressions 66 to 69 from the proof of Lemma 5, with $w = 0$:

$$\forall x : Pr_{\sigma \sim P} \{c_R'''(x)\} \leq \left(\frac{(k+r-1)m}{n} \right)^r e^r. \quad (102)$$

So

$$E_{(x,y) \sim D} \{Pr_{\sigma \sim P} \{c_R'''(x, \sigma)\}\} \leq \left(\frac{(k+r-1)m}{n} \right)^r e^r, \quad (103)$$

and, combining this with Expressions 99 and 100:

$$|E_{\sigma \sim P} \{t_W(\sigma)\}| \leq 2^{r-1} \left(\frac{(k+r-1)m}{n} \right)^r e^r. \quad (104)$$

Substitute this into Inequality 95:

$$Pr_{F \sim D^n} \left\{ |E_Q \{t_W(\sigma)\}| \geq 2^r \left[\sqrt{\frac{\ln \frac{2}{\delta_q}}{2q}} + \left(\frac{(k+r-1)m}{n} \right)^r e^r \right] \right\} \leq \delta_q. \quad (105)$$

Combine the bound on $E_Q \{t_V(\sigma)\}$ in Inequality 91 with this bound on $E_Q \{t_W(\sigma)\}$ to produce a bound for p^* :

$$\begin{aligned} Pr_{F \sim D^n} \left\{ |p^* - E_Q \{s_V(\sigma)\}| \geq r 3^{r-1} \sqrt{\frac{\ln \frac{2rq}{\delta}}{2m}} + 2^r \left[\sqrt{\frac{\ln \frac{2}{\delta_q}}{2q}} + \left(\frac{(k+r-1)m}{n} \right)^r e^r \right] \right\} \\ \leq \delta + \delta_q. \end{aligned} \quad \begin{matrix} (106) \\ (107) \end{matrix}$$

■

Proof [of Corollary 7] Recall that

$$\epsilon_q = r 3^{r-1} \sqrt{\frac{\ln \frac{2rq}{\delta}}{2m}} + 2^r \left[\sqrt{\frac{\ln \frac{2}{\delta_q}}{2q}} + \left(\frac{(k+r-1)m}{n} \right)^r e^r \right]. \quad (108)$$

Select

$$m = \frac{n^{\frac{r}{r+1/2}}}{(k+r-1)e} \quad (109)$$

and $q = n$. Then

$$\epsilon_q = n^{-\frac{r}{2r+1}} \left[r 3^{r-1} \sqrt{\frac{1}{2}(k+r-1)e \ln \frac{2rn}{\delta}} + 1 \right] + 2^r \sqrt{\frac{\ln \frac{2}{\delta_q}}{2n}}. \quad (110)$$

Holding r constant, this is $O\left(n^{-\frac{r}{2r+1}} \sqrt{k \ln n}\right)$. ■

Proof [of Corollary 8] The proof of this Corollary is very similar to the proof of Corollary 3. Let $r = \lceil C\sqrt{\ln n} \rceil$. Then, based on Equation 110,

$$\epsilon_q = O\left(n^{-\frac{\lceil C\sqrt{\ln n} \rceil}{2\lceil C\sqrt{\ln n} \rceil + 1}} 3^{C\sqrt{\ln n}} \sqrt{k \ln n}\right). \quad (111)$$

As in the proof of Corollary 3, convert the exponent on 3 to an exponent on n , and let $C = \frac{1}{2\sqrt{\ln 3}}$. Then

$$\epsilon_q = O\left(n^{-\frac{1}{2} + \sqrt{\frac{\ln 3}{\ln n}}} \sqrt{k} \ln n\right). \quad (112)$$

■

We can improve Theorem 6 by directly computing

$$E_{\sigma \sim P} \{Pr_{(x,y) \sim D} \{c_R(x)\}\} \quad (113)$$

from the RHS of Inequality 97 rather than using $Pr_{(x,y) \sim D} \{c_R'''(x)\}$ to bound $Pr_{(x,y) \sim D} \{c_R(x)\}$. This generates a tighter error bound, but it is more difficult to prove asymptotic results about this tighter bound:

Theorem 9 For $\delta > 0$ and $\delta_q > 0$:

$$Pr_{F \sim D^n} \{|p^* - E_Q \{s_V(\sigma)\}| \geq \epsilon_q\} \leq \delta + \delta_q, \quad (114)$$

where

$$\epsilon_q = r3^{r-1} \sqrt{\frac{\ln \frac{2rq}{\delta}}{2m}} + 2^r \left[\sqrt{\frac{\ln \frac{2}{\delta_q}}{2q}} + u(n, k, r) \right], \quad (115)$$

and

$$u(n, k, r) = \sum_{i=r}^{n-k} \binom{n}{k+i}^{-1} \binom{n-rm}{k} \frac{k}{k+i} \left[\sum_{j=0}^r (-1)^j \binom{r}{j} \binom{(r-j)m}{i} \right]. \quad (116)$$

Proof [of Theorem 9] In the proof of Theorem 6, in Inequality 102, we used the RHS to bound $Pr_{\sigma \sim P} \{c_R'''(x)\}$, which is a bound for $Pr_{\sigma \sim P} \{c_R(x)\}$. So we can replace

$$\left(\frac{(k+r-1)m}{n} \right)^r e^r \quad (117)$$

in the statement of Theorem 6 by $Pr_{\sigma \sim P} \{c_R(x)\}$. (It is the same for all x .) So we need to show that $Pr_{\sigma \sim P} \{c_R(x)\} = u(n, k, r)$.

Define some conditions on x and σ :

1. Let u_i' be the condition that the nearest $k+i$ examples to x from σF include k from $\sigma F - V(\sigma)$ and i from $V(\sigma)$.
2. Let u_i'' be true if and only if the $(k+i)$ th nearest example to x from σF is from $\sigma F - V(\sigma)$.
3. Let u_i''' be the condition that the nearest i examples to x from $V(\sigma)$ include at least one example from each of $V_1(\sigma), \dots, V_r(\sigma)$.

Note that

$$Pr_{\sigma \sim P} \{c_R(x)\} = \sum_{i=r}^{n-k} Pr_{\sigma \sim P} \{u'_i\} Pr_{\sigma \sim P} \{u''_i | u'_i\} Pr_{\sigma \sim P} \{u'''_i\}, \quad (118)$$

$$Pr_{\sigma \sim P} \{u'_i\} = \binom{n}{k+i}^{-1} \binom{n-rm}{k} \binom{rm}{i}, \quad (119)$$

and

$$Pr_{\sigma \sim P} \{u''_i | u'_i\} = \frac{k}{k+i}. \quad (120)$$

Let z_S be the condition that the nearest i examples to x from $V(\sigma)$ have no examples from any $V_i(\sigma)$ indexed by $S \subseteq R$. By inclusion and exclusion:

$$Pr_{\sigma \sim P} \{u'''_i\} = 1 - r Pr_{\sigma \sim P} \{z_{\{1\}}\} + \binom{r}{2} Pr_{\sigma \sim P} \{z_{\{1,2\}}\} - \dots \pm \binom{r}{r} Pr_{\sigma \sim P} \{z_R\}. \quad (121)$$

Gathering terms and filling in values for $Pr_{\sigma \sim P} \{z_S\}$:

$$Pr_{\sigma \sim P} \{u'''_i\} = \sum_{j=0}^r (-1)^j \frac{\binom{r}{j} \binom{(r-j)m}{i}}{\binom{rm}{i}} \quad (122)$$

Substitute Equations 119, 120, and 122 into Equation 118, and cancel $\binom{rm}{i}$. ■

5. Discussion

We have shown that k -nn classifiers have data-independent error bounds with $O\left(n^{-\frac{r}{2r+1}} \sqrt{k \ln n}\right)$ error bound ranges. This has a form similar to classical VC-style bounds (Vapnik and Chervonenkis, 1971) that have bound range $O\left(n^{-\frac{1}{2}} \sqrt{d \ln n}\right)$ for classifiers selected from hypothesis classes with VC dimension d . In the future, it would be interesting to extend the k -nn error bounds to cover selection of a distance metric from a parameterized set of “hypothesis” metrics (Kedem et al., 2012).

It may be possible to improve the data-independent bounds in this paper by averaging bounds over all permutations of the in-sample data, or, equivalently, over all possible choices of V_1, \dots, V_r . In practice, this should reduce bias. The theoretical challenge is to show that the average over bounds for different permutations does not have a larger bound range than a bound over a single permutation. For some theory on averaging bounds to form a single bound, refer to Bax (1998), McAllester (1999), and Langford et al. (2001). The practical challenge is to efficiently compute a bound over many permutations of the in-sample data. For a method to solve a similar problem for leave-one-out estimates, refer to Mullin and Sukthankar (2000).

It may be possible to improve on the results in this paper by tightening some bounds used to derive the error bounds. In the random variable corresponding to a validation example, it may not be logically possible for all positive terms ($|T|$ even) to be nonzero while the negative terms ($|T|$ odd) are zero, or vice versa. So the range of these variables may be much less than the equations indicate, and perhaps not even exponential in r . Also, it may be possible to improve the bound in Lemma 5

by using the requirement of $c'_R(x)$ that *all* of V_1, \dots, V_r contribute nearer neighbors to x than the k^{th} nearest neighbor from $(F - V) - W$. In addition, it may be possible to derive good analytic bounds for the combinatorial expression for $Pr_{\sigma \sim P} \{c_R(x)\}$ derived in the proof of Theorem 9. This may allow larger validation set sizes, producing bounds better than those in Corollaries 7 and 8.

Finally, it would be interesting to apply the techniques from this paper to derive error bounds for network classifiers, where the data is a graph annotated with node and edge data, and the goal is to generalize from labels on some nodes to labels for unlabeled nodes, sometimes including nodes yet to be added to the graph. (See Sen et al. (2008) and Macskassy and Provost (2007) for more background on collective classification.) An initial challenge is to adapt the methods in this paper to network settings where the classification rules are local – based only on neighbors or neighbors of neighbors in the graph – and where nodes are drawn i.i.d. In this setting, nodes are similar to examples, and neighborhoods in the graph have a role similar to near neighbor relationships in the k -nn setting. It will be more challenging to apply the techniques to settings where classification rules are not local or nodes are not drawn i.i.d. For some background on error bounds in such settings, refer to London et al. (2012), Li et al. (2012) and Bax et al. (2013).

References

- J.-Y. Audibert. *PAC-Bayesian Statistical Learning Theory*. PhD thesis, Laboratoire de Probabilités et Modèles Aléatoires, Universités Paris 6 and Paris 7, 2004. URL <http://cermis.enpc.fr/~{ }audibert/ThesePack.zip>.
- J.-Y. Audibert, R. Munos, and Csaba Szepesvari. Variance estimates and exploration function in multi-armed bandit. *CERTIS Research Report 07-31*, 2007.
- E. Bax. Validation of average error rate over classifiers. *Pattern Recognition Letters*, pages 127–132, 1998.
- E. Bax. Nearly uniform validation improves compression-based error bounds. *Journal of Machine Learning Research*, 9:1741–1755, 2008.
- E. Bax. Validation of k -nearest neighbor classifiers. *IEEE Transactions on Information Theory*, 58(5):3225–3234, 2012.
- E. Bax and A. Callejas. An error bound based on a worst likely assignment. *Journal of Machine Learning Research*, 9:581–613, 2008.
- E. Bax and Y. Le. Some theory for practical classifier validation. *Baylearn*, 2015. URL <http://arxiv.org/abs/1510.02676>.
- E. Bax, J. Li, A. Sonmez, and Z. Cataltepe. Validating collective classification using cohorts. *NIPS Workshop on Frontiers of Network Analysis: Methods, Models, and Applications*, 2013. URL http://snap.stanford.edu/networks2013/papers/netnips2013_submission_11.pdf.
- S. N. Bernstein. On certain modifications of Chebyshev’s inequality. *Doklady Akademii Nauk SSSR*, 17(6):275–277, 1937.

- A. Blum and J. Langford. PAC-MDL bounds. In *Proceedings of the 16th Annual Conference on Computational Learning Theory (COLT)*, pages 344–357, 2003.
- S. Boucheron, G. Lugosi, and P. Massart. *Concentration Inequalities – A Nonasymptotic Theory of Independence*. Oxford University Press, 2013.
- V. Chvátal. The tail of the hypergeometric distribution. *Discrete Mathematics*, 25(3):285–287, 1979.
- T. M. Cover. Rates of convergence for nearest neighbors procedures. In B. K. Kinariwala and F. F. Kuo, editors, *Proceedings of the Hawaii International Conference on System Sciences*, pages 413–415. University of Hawaii Press, 1968.
- T. M. Cover and P. E. Hart. Nearest neighbor pattern classification. *IEEE Transactions on Information Theory*, 13:21–27, 1967.
- L. Devroye and T. Wagner. Distribution-free inequalities for the deleted and holdout estimates. *IEEE Transactions on Information Theory*, 25:202–207, 1979.
- L. Devroye, L. Györfi, and G. Lugosi. *A Probabilistic Theory of Pattern Recognition*. Springer, 1996.
- R. O. Duda, P. E. Hart, and D. G. Stork. *Pattern Classification*. Wiley, 2001.
- S. Floyd and M. Warmuth. Sample compression, learnability, and the Vapnik-Chervonenkis dimension. *Machine Learning*, 21(3):1–36, 1995.
- T. Hastie, R. Tibshirani, and J. Friedman. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction, Second Edition*. Springer, 2009.
- W. Hoeffding. Probability inequalities for sums of bounded random variables. *Journal of the American Statistical Association*, 58(301):13–30, 1963.
- P. G. Hoel. *Introduction to Mathematical Statistics*. Wiley, 1954.
- Dor Kedem, Stephen Tyree, Fei Sha, Gert R. Lanckriet, and Kilian Q Weinberger. Non-linear metric learning. In F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 25*, pages 2573–2581. Curran Associates, Inc., 2012. URL <http://papers.nips.cc/paper/4840-non-linear-metric-learning.pdf>.
- J. Langford. Tutorial on practical prediction theory for classification. *Journal of Machine Learning Research*, 6:273–306, 2005.
- John Langford, Matthias Seeger, and Nimrod Megiddo. An improved predictive accuracy bound for averaging classifiers. In *In Proceeding of the Eighteenth International Conference on Machine Learning*, pages 290–297, 2001.
- J. Li, A. Sonmez, Z. Cataltepe, and E. Bax. Validation of network classifiers. *Structural, Syntactic, and Statistical Pattern Recognition Lecture Notes in Computer Science*, 7626:448–457, 2012.

- N. Linial and N. Nisan. Approximate inclusion-exclusion. *Combinatorica*, 10(4):349–365, 1990.
- N. Littlestone and M. Warmuth. Relating data compression and learnability, 1986. Unpublished manuscript, University of California Santa Cruz.
- Ben London, Bert Huang, and Lise Getoor. Improved generalization bounds for large-scale structured prediction. In *NIPS Workshop on Algorithmic and Statistical Approaches for Large Social Networks*, 2012.
- S. A. Macskassy and F. Provost. Classification in networked data: A toolkit and a univariate case study. *Journal of Machine Learning Research*, 8:935–983, 2007.
- Andreas Maurer and Massimiliano Pontil. Empirical bernstein bounds and sample-variance penalization. *22nd Annual Conference on Learning Theory (COLT)*, 2009.
- David A. McAllester. Pac-bayesian model averaging. In *In Proceedings of the Twelfth Annual Conference on Computational Learning Theory*, pages 164–170. ACM Press, 1999.
- V. Mnih, C. Szepesvari, and J.-Y. Audibert. Empirical Bernstein stopping. *Proceedings of the 25th International Conference on Machine Learning*, pages 672–679, 2008.
- M. Mullin and R. Sukthankar. Complete cross-validation for nearest neighbor classifiers. *Proceedings of the Seventeenth International Conference on Machine Learning*, pages 639–646, 2000.
- D. Psaltis, R. Snapp, and S. Venkatesh. On the finite sample performance of the nearest neighbor classifier. *IEEE Transactions on Information Theory*, 40(3):264–280, 1994.
- P. Sen, G. Namata, M. Bilgic, L. Getoor, B. Gallagher, and T. Eliassi-Rad. Collective classification in network data. *AI Magazine*, 29(3):93–106, 2008.
- L. G. Valiant. A theory of the learnable. *Commun. ACM*, 27(11):1134–1142, 1984. ISSN 0001-0782. doi: <http://doi.acm.org/10.1145/1968.1972>.
- V. Vapnik and A. Chervonenkis. On the uniform convergence of relative frequencies of events to their probabilities. *Theory of Probability and its Applications*, 16:264–280, 1971.

Appendix A. Tests

This section presents test results to show that using $r > 2$ can improve error bounds even for medium-sized data sets. We start with some modifications to make the results from the previous section produce stronger bounds. Then we present test results.

The random variables corresponding to validation examples in t_v tend to have absolute values that are small compared to their ranges: $c_{S \cup T}(x)$ becomes more unlikely as $|S \cup T|$ grows, and $g_S(x) \neq y$ is rare for accurate classifiers, making $I((x, y) \in C_{S,T})$ zero in many cases. So to make the bounds stronger, we use empirical Bernstein bounds in place of Hoeffding bounds for sums of those random variables. Empirical Bernstein bounds were first developed by Audibert (Audibert, 2004; Audibert et al., 2007; Mnih et al., 2008) and are based on Bernstein bounds (Bernstein, 1937). We use the version of empirical Bernstein bounds by Maurer and Pontil (2009). (For a variety of

similar bounds, refer to Boucheron et al. (2013).) Empirical Bernstein bounds are stronger than the standard Hoeffding bounds when the random variables have small standard deviations compared to their ranges. In effect, empirical Bernstein bounds bound the variance of the random variable, then rely on a small variance to produce a strong bound on the mean. (Hoeffding (1963) includes a strong bound for low-variance random variables, but the standard version of Hoeffding bounds is based on a worst-case assumption about the variance.)

As r increases, the random variables in $Pr_W \{c'_R(x)\}$ become increasingly likely to be zeroes. Since these random variables have value either zero or one, we use directly computed binomial tail bounds, as described by Langford (2005) and Hoel (1954) (page 208). When $Pr_{(x,y) \sim D} \{c'_R(x)\}$ is near zero, these bounds take advantage of the low variance in $Pr_W \{c'_R(x)\}$ terms.

To further improve the bounds, we truncate the inclusion and exclusion formulas:

$$Pr_{(x,y) \sim D} \{B_S\} = \sum_{T \subseteq R-S} (-1)^{|T|} Pr_{(x,y) \sim D} \{C_{S,T}\}. \quad (123)$$

To truncate, select an even $u \geq 0$, and

$$Pr_{(x,y) \sim D} \{B_S\} \leq \sum_{T \subseteq R-S, |T| \leq u} (-1)^{|T|} Pr_{(x,y) \sim D} \{C_{S,T}\}. \quad (124)$$

(For more on truncation of inclusion and exclusion, refer to Linial and Nisan (1990).) For a depth parameter $0 \leq d \leq r$, let $u(S) = \max(2 \lfloor \frac{d-|S|}{2} \rfloor, 0)$. Then a truncated version of Theorem 4 is

$$Pr_{(x,y) \sim D} \{g^*(x) \neq y\} \leq \sum_{S \subseteq R} \sum_{T \subseteq R-S, |T| \leq u(S)} (-1)^{|T|} Pr_{(x,y) \sim D} \{C_{S,T}\}. \quad (125)$$

When we truncate to depth d , we estimate only the terms in this truncated formula, ignoring the others. This decreases the range of the random variables that correspond to validation examples, and it decreases the number of terms with $|S \cup T| = |R|$ that we bound based on $Pr_W \{c'_R(x)\}$. In fact, for $d < r$, there is only one such term: $Pr_{(x,y) \sim D} \{c_R(x) \wedge (g_R(x) \neq y)\}$. So, in t_w , the bound on $Pr_{(x,y) \sim D} \{c'_R(x)\}$ is multiplied by one instead of a coefficient that is exponential in r .

Here is pseudocode for a one-sided (upper) bound, incorporating empirical Bernstein bounds, directly computed binomial tail bounds, and truncated inclusion and exclusion:

testBound

1. inputs: data set F , $r > 0$, $|V_1|, \dots, |V_r|$, $\delta > 0$, $|W|$, $\delta_W > 0$, $d \in \{0, \dots, r\}$
2. sum = 0.0.
3. // Bound t_V :
4. Randomly partition: $F \rightarrow (F - V, V_1, \dots, V_r)$.
5. for $i \in \{1, \dots, r\}$:

- (a) range = $|V_i| \sum_{S \subseteq R - \{i\}} \sum_{T \subseteq (R - \{i\}) - S, |T| \leq \max(2 \lfloor \frac{d-|S|}{2} \rfloor, 0)} \frac{1}{|V_{R-(S \cup T)}|}$.
- (b) values = $(\forall (x, y) \in V_i :$
- (c) $|V_i| \left[\sum_{S \subseteq R - \{i\}} \sum_{T \subseteq (R - \{i\}) - S, |T| \leq \max(2 \lfloor \frac{d-|S|}{2} \rfloor, 0)} (-1)^{|T|} \frac{1}{|V_{R-(S \cup T)}|} I((x, y) \in C_{S,T}) \right]$).

- (d) $\text{sum} = \text{sum} + \text{empBernsteinBound}(\text{values}, \text{range}, \frac{\delta}{r})$.
6. // Bound t_W :
7. Randomly partition: $F - V \rightarrow (F - V - W, W)$.
8. $\text{values} = (\forall (x, y) \in W : c'_R(x))$.
9. $\text{sum} = \text{sum} + \text{directBound}(\text{values}, \delta_W)$.
10. return sum .

empBernsteinBound

1. inputs: values , range , $\delta > 0$
2. return $\text{mean}(\text{values}) + \sqrt{\frac{2\text{Var}(\text{values}) \ln \frac{2}{\delta}}{|\text{values}|}} + \text{range} \frac{7 \ln \frac{2}{\delta}}{3(|\text{values}| - 1)}$.

(For directBound , refer to Langford (2005) or Hoel (1954), page 208.)

We ran tests for $1 \leq r \leq 5$ and depth $0 \leq d < r$, for $k = 3$ and $k = 7$, with $n = 50,000$. For each test, we generated n in-sample examples at random, with x drawn uniformly from a bounded cube centered at the origin. Each label y depends on whether the number of negative components in x is even or odd. If it is even, then the label is one with probability 90% and zero with probability 10%. If it is odd, then the probabilities are reversed. (So the label depends on the quadrant, with 10% of labels flipped to add some noise.)

For each test, we applied the in-sample examples as a k -nn classifier to 100,000 random out-of-sample examples to estimate the expected out-of-sample error rate. For each r and

$$c \in \{0.625\%, 1.25\%, 2.5\%, 3.75\%, \dots, 10\%\}, \quad (126)$$

we randomly partitioned the examples into $F - V - W, V_1, \dots, V_r, W$ with $m = |V_i| = cn$ (rounded to the nearest integer), and $|W| = |V_i|$. Then we computed an upper bound on expected out-of-sample error rate using each truncation depth $0 \leq d < r$. We recorded differences between bounds and performance on the 100,000 out-of-sample examples. Out-of-sample error rates were about 15% for $k = 3$ and about 12% for $k = 7$. Each result is the average of 100 tests, and the standard deviations for differences between bounds and estimated out-of-sample error rates are about 1%, making the standard deviations of the estimates of the means over the 100 trials about 0.1%. (So, in the figures, the differences between plotted points are statistically significant.)

Figures 5 and 6 show results for $k = 3$ and $k = 7$, respectively. For each r value, the figures show the curve for the d value that yields the tightest bound. For $k = 3$, the smallest average difference between bounds and estimated out-of-sample error rates (over 100 trials) was 1.3%, achieved with $r = 3, d = 2$, and $m = 6.25\%$ of n . So $3 \cdot 6.25\%$, or about 20%, of the in-sample examples were used for validation (or 25% if we count examples in W as well as V). For $k = 7$, the smallest average difference between bounds and estimated out-of-sample error rates is 10.5%, achieved with $r = 3, d = 2$, and $m = 3.75\%$ of n . As k increases from 3 to 7, $Pr_{(x,y) \sim D} \{c'_R(x)\}$ increases; decreasing validation set sizes helps offset the increase. (In Figure 6, the tightest bound for $r = 1$ and $d = 0$ lies just beyond the left side of the figure: it is 14.5%, achieved at $m = 0.6\%$ of n .) For $k = 3$ and $n = 25,000$ (not shown in figures), the minimum average gap between bound and test

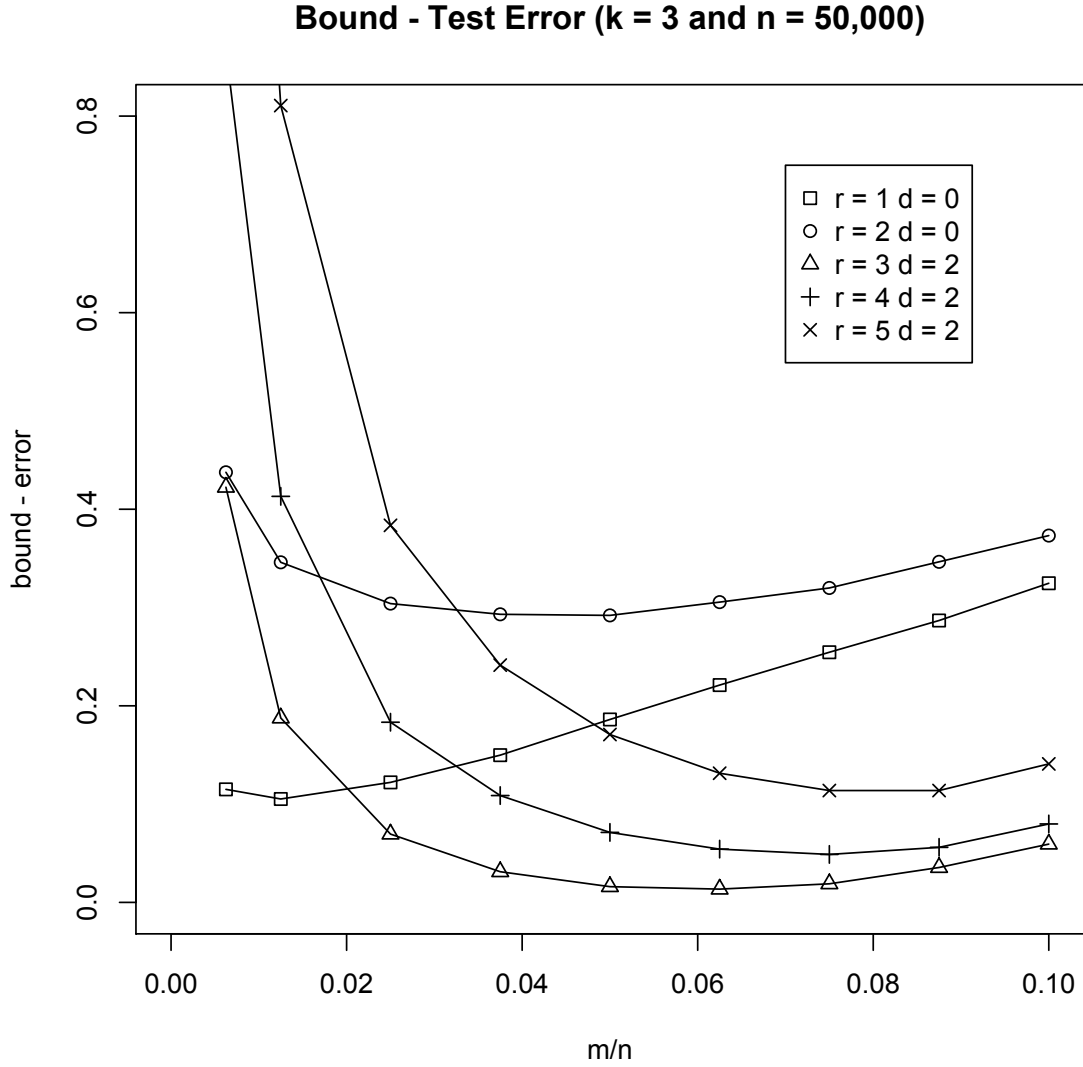


Figure 5: Differences between upper bound on out-of-sample error rate and actual error rate over 100,000 out-of-sample examples, averaged over 100 tests. The tightest bound is achieved with $r = 3$, $d = 2$, and $m = 3125$ (which is 6.25% of n).

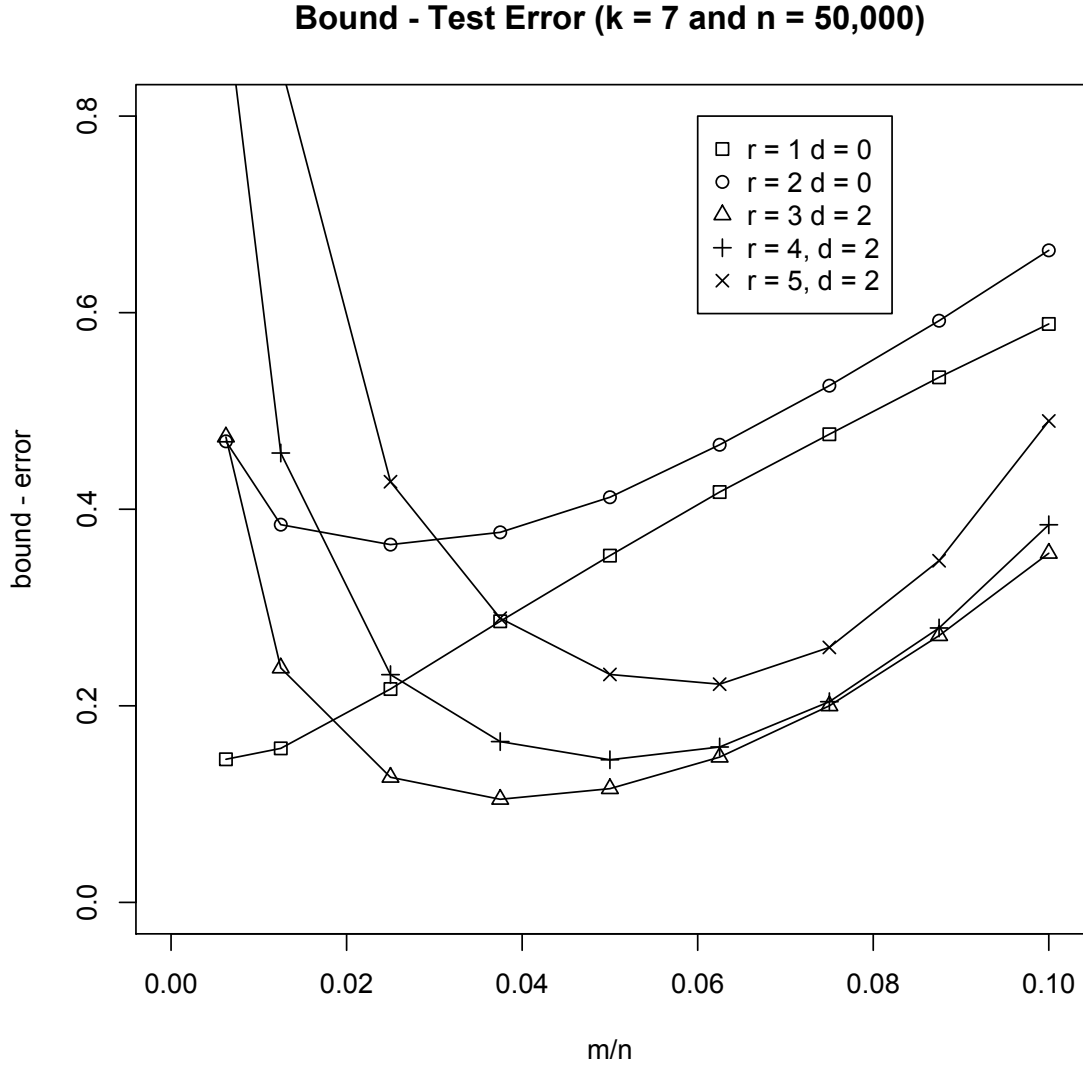


Figure 6: Differences between upper bound on out-of-sample error rate and actual error rate over 100,000 out-of-sample examples, averaged over 100 tests. The tightest bound is achieved with $r = 3$, $d = 2$, and $m = 1875$ (which is 3.75% of n).

error was 6%, also achieved with $r = 3$ and $d = 2$, but with $m = 7.5\%$ of n , indicating that as the number of in-sample examples shrinks, a larger fraction of them are needed for validation.

Choices of r and d mediate tradeoffs in bound tightness. From Equation 62, with $d = r$, $Pr_W \{c'_R(x)\}$ tends to shrink exponentially in r , since $c'_R(x)$ requires all r validation sets to have neighbors near x . However, the coefficients $r3^{r-1}$ and 2^r increase exponentially with r . (There is also an increase proportional to $\sqrt{\ln r}$, because we bound over more validation sets.) But Equation 62 is an upper bound on the difference between the error bound and the out-of-sample error rate. In practice, we can replace 3^{r-1} by the length of the range of each $f_i(x, y)$:

$$\sum_{S \subseteq R - \{i\}} \sum_{T \subseteq (R - \{i\}) - S} \frac{|V_i|}{|V_{R-(S \cup T)}|} = \sum_{S \subseteq R - \{i\}} \sum_{T \subseteq (R - \{i\}) - S} \frac{1}{|R - (S \cup T)|}. \quad (127)$$

Using empirical Bernstein (vs. Hoeffding) bounds reduces the impact of this range on the bound. Using truncated inclusion and exclusion ($d < r$) reduces the range by removing terms from the double sum and also by reducing the coefficient on $Pr_W \{c'_R(x)\}$ from 2^r to one. The tradeoff is that truncation introduces a bias into the bound. However, the truncated terms tend to be small compared to the remaining terms, because the truncated terms require more validation sets to have neighbors near an example, due to condition $c_{S \cup T}(x)$.

Appendix B. A Separate Validation per Combination of Subsets

In the main text, we do r separate validations, one for each validation subset V_i , then we use a sum bound on a probability of a union to form a bound on $|s_V - t_V|$. In this appendix, we show that, alternatively, we may do a separate validation for each combination of validation subsets and get a similar result. Recall (from Equality 39) that

$$t_V = \sum_{S \subset R} \sum_{T \subset R - S} (-1)^{|T|} Pr_{(x,y) \sim D} \{C_{S,T}\}. \quad (128)$$

Let $A = S \cup T$, and rewrite t_V as a sum over A :

$$t_V = \sum_{A \subset R} \sum_{S \subseteq A} (-1)^{|A-S|} Pr_{(x,y) \sim D} \{C_{S,A-S}\}. \quad (129)$$

Define

$$f_A(x, y) \equiv \sum_{S \subseteq A} (-1)^{|A-S|} I((x, y) \in C_{S,A-S}). \quad (130)$$

Then

$$t_V = \sum_{A \subset R} E_{(x,y) \sim D} \{f_A(x, y)\}. \quad (131)$$

Since

$$\forall (x, y) : f_A(x, y) | F = f_A(x, y) | F - V_{R-A}, \quad (132)$$

we can apply the Hoeffding Inequality to each $f_A(x, y)$, using an empirical mean over V_{R-A} . So,

$$\forall A \subset R : Pr_{F \sim D^n} \left\{ \left| E_{V_{R-A}} \{f_A(x, y)\} - E_{(x,y) \sim D} \{f_A(x, y)\} \right| \geq 2^{|A|} \sqrt{\frac{\ln \frac{2}{\delta_A}}{2|V_{R-A}|}} \right\} \leq \delta_A, \quad (133)$$

where $\delta_A > 0$ for all A , and the probability is over random draws of F . Using a sum bound for the probability of a union,

$$Pr_{(x,y) \sim D} \left\{ \left| \sum_{A \subset R} E_{V_{R-A}} \{f_A(x, y)\} - \sum_{A \subset R} E_{(x,y) \sim D} \{f_A(x, y)\} \right| \geq \sum_{A \subset R} 2^{|A|} \sqrt{\frac{\ln \frac{2}{\delta_A}}{2|V_{R-A}|}} \right\} \leq \sum_{A \subset R} \delta_A. \quad (134)$$

The first sum in the absolute value is s_V , and the second sum is t_V . Define

$$\epsilon_V \equiv \sum_{A \subset R} 2^{|A|} \sqrt{\frac{\ln \frac{2}{\delta_A}}{2|V_{R-A}|}} \quad (135)$$

and let

$$\delta = \sum_{A \subset R} \delta_A. \quad (136)$$

Then

$$Pr_{F \sim D^n} \{|s_V - t_V| \geq \epsilon_V\} \leq \delta. \quad (137)$$

Since the $f_A(x, y)$ with different $|A|$ values have different ranges and different-sized validation sets V_{R-A} , it is not optimal to set all δ_A to the same value. To optimize, we could take the partial derivatives of ϵ_V with respect to each δ_A , set those partial derivatives equal to each other, and solve (numerically) for the optimal δ_A values under the constraint that they sum to δ . (To simplify this optimization, note that, by symmetry, it is optimal to set all δ_A with the same $|A|$ to the same value.)

For a straightforward result, let δ_j be the value of each δ_A having $|A| = j$. Then

$$\epsilon_V = \sum_{j=0}^{r-1} \binom{r}{j} 2^j \sqrt{\frac{\ln \frac{2}{\delta_j}}{(r-j)m}}, \quad (138)$$

and

$$\delta = \sum_{j=0}^{r-1} \binom{r}{j} \delta_j. \quad (139)$$

Set

$$\delta_j = 2 \left(\frac{\delta}{2r\alpha(\delta)} \right)^{r-j}, \quad (140)$$

where

$$\alpha(\delta) \equiv \frac{\frac{\delta}{2}}{\ln(1 + \frac{\delta}{2})}. \quad (141)$$

(Note that $\alpha(\delta)$ is close to one, since $z \approx \ln(1 + z)$ is a well-known approximation for small z .)

Then

$$\epsilon_V = \sum_{j=0}^{r-1} \binom{r}{j} 2^j \sqrt{\frac{\ln \frac{2r\alpha(\delta)}{\delta}}{m}}. \quad (142)$$

By binomial expansion,

$$\sum_{j=0}^r \binom{r}{j} 2^j = (1 + 2)^r = 3^r, \quad (143)$$

so

$$\sum_{j=0}^{r-1} \binom{r}{j} 2^j = 3^r - 2^r. \quad (144)$$

So

$$\epsilon_V = (3^r - 2^r) \sqrt{\frac{\ln \frac{2r\alpha(\delta)}{\delta}}{m}}. \quad (145)$$

To show that this bound is valid, we need to show that

$$\sum_{j=0}^{r-1} \binom{r}{j} \delta_j \leq \delta. \quad (146)$$

We can do this as follows:

$$\sum_{j=0}^{r-1} \binom{r}{j} \delta_j \quad (147)$$

$$= \sum_{j=0}^{r-1} \binom{r}{j} 2 \left(\frac{\delta}{2r\alpha(\delta)} \right)^{r-j} \quad (148)$$

$$= 2 \left(-1 + \sum_{j=0}^r \binom{r}{j} \left(\frac{\delta}{2r\alpha(\delta)} \right)^{r-j} \right) \quad (149)$$

$$2 \left(-1 + \left(1 + \frac{\delta}{2r\alpha(\delta)} \right)^r \right) \quad (150)$$

$$\leq 2 \left(-1 + e^{\frac{\delta}{2} \frac{1}{\alpha(\delta)}} \right) \quad (151)$$

$$= 2 \left(-1 + e^{\ln(1+\frac{\delta}{2})} \right) \quad (152)$$

$$= \delta. \quad (153)$$